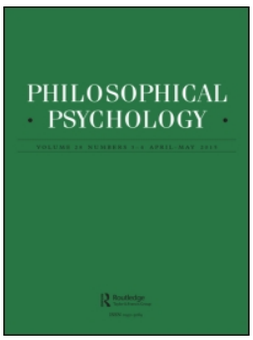

Routledge
Taylor & Francis Group

Philosophical Psychology

ISSN: 0951-5089 (Print) 1465-394X (Online) Journal homepage: https://www.tandfonline.com/loi/cphp20

Implicit attitudes and the social capacity for free will

Daphne Brandenburg

To cite this article: Daphne Brandenburg (2016) Implicit attitudes and the social capacity for free will, Philosophical Psychology, 29:8, 1215-1228, DOI: 10.1080/09515089.2016.1235263

To link to this article: https://doi.org/10.1080/09515089.2016.1235263

PHILOSOPHICAL PSYCHOLOGY, 2016
VOL. 29, NO. 8, 1215–1228
http://dx.doi.org/10.1080/09515089.2016.1235263

Routledge
Taylor & Francis Group

Implicit attitudes and the social capacity for free will

Daphne Brandenburg

Department of Philosophy, Theology, and Religious Studies, Radboud University, Nijmegen, The Netherlands

ABSTRACT

In this paper I ask what implicit attitudes tell us about our freedom. I analyze the relation between the literature on implicit attitudes and an important subcategory of theories of free will—self-disclosure accounts. If one is committed to such a theory, I suggest one may have to move to a more social conceptualization of the capacity for freedom. I will work out this argument in five sections. In the first section, I discuss the specific theories of free will that are central to this paper. In the second section, I will show that implicit-bias research raises questions about people's capacities to exercise (these specific understandings of) free will. In the third section, I will consider how an individual may overcome these failures and argue that the individual ability for self-regulation is significantly limited. One could stop here and conclude that free will is a limited capacity. But I argue that this conclusion would be too hastily drawn. I will instead continue to ask what would be required for free will. By discussing how failures of free will are due to social structures and may be therefore repaired by changing social structures in section 4, I will arrive at an alternative conclusion about the capacity for free will in section 5.

ARTICLE HISTORY
Received 17 April 2016
Accepted 29 August 2016

KEYWORDS
Associative learning; free will; implicit attitudes; relational autonomy; self-disclosure; self-regulation; social scaffolding

> If he's Black, unless he has a big smile on his face, then I become mildly racist and think, "That's fine, everything's fine, nothing's going to happen." Of course I'm fine. Why did I even think that for a second?
>
> Louis C. K., 16 May 2015

On *Saturday Night Live*, Louis C. K. jokingly commented on the undesirable implicit attitudes of distrust or suspicion he sometimes feels toward African Americans. To readers, the form of "mild racism" he describes is probably familiar. It is commonly argued that almost everyone in the western world harbors prejudiced implicit attitudes toward stigmatized groups. In this paper, I ask what implicit attitudes tell us about our freedom. I analyze the relation between the literature on implicit attitudes and an important subcategory of theories of free will—self-disclosure accounts. If one is committed to such a theory, I suggest that one may have to move to a more social conceptualization of the capacities for freedom.

I will work out this argument in five sections. In the first section, I discuss the specific theories of free will that are central to this paper. In the second section, I will show that research on implicit attitudes raises questions about people's capacities to exercise (these specific understandings of) free will. This is because research on implicit attitudes highlights failures of free will on these accounts. In the third section, I will contemplate how one may overcome these failures and argue that the individual

CONTACT Daphne Brandenburg daphnebrandenburg@gmail.com

ability for self-regulation is significantly limited. One could stop here and conclude that free will is a limited capacity. But I argue that this conclusion would be too hastily drawn. I will instead continue to ask what would be required for free will. By discussing how failures of free will are due to social structures and may therefore be repaired by changing social structures in section 4, I will arrive at an alternative conclusion in section 5. I will conclude that the capacity to exercise free will should be conceptualized as partly social and will situate this conclusion in the existing literature on social re-conceptualizations of freedom and autonomy.

1. Freedom

The notions of free will that are at stake in this paper are "self-disclosure" accounts.[1] Such accounts propose a form of wholeheartedness, identification, or self-expression as the mark of free action (e.g., Bratman, 2007; Christman, 2009; Frankfurt, 1988; Sripada, 2015; Watson, 1975). On these accounts, one has free will when one's motives and desires align with one's authentic commitments and cares or can be resisted when they diverge from them. Typical examples of a failure to act on the basis of authentic commitments would be actions that result from phobias or instances of compulsive action from which the agent feels alienated and would like to be rid of. Examples of inauthentic motives or desires, on the less extreme end of the spectrum, include the desire to eat a slice of cake even though one is seriously committed to dieting or pangs of jealousy when one's partner is talking to someone at a party although one does not regard this jealousy as justified or appropriate in the light of the relationship and values that one has. What is common to all of these examples is that these motives or desires would give rise to actions that one would not endorse or accept in the light of one's commitments, plans, and cares, and are as such motives or desires that one would rather not act upon.

Surely there is no one who *always* acts on the basis of motives or desires that align with his or her authentic normative outlook on life, nor is it likely that one is always fully able to do so. On these conceptions of free will, no human being can really be said to be absolutely and totally free. But it is generally assumed that people have some ability to exercise the relevant form of freedom. And this ability to act in a way that is consistent with one's authentic commitments and cares is the conception of freedom that will be central to this paper.

Admittedly, this conception is tentative and unspecified as it stands. One of the main challenges for such accounts is to explain what it means to form and have *authentic* cares and commitments. This is not an easy task, and it is a task that is far beyond the scope of this paper. Luckily, my argument does not require a specification of the general conditions for authenticity. All that is needed for my argument are a number of concrete cases of which it can be said that people really *fail* to act on the basis of their authentic cares and commitments. I discuss these cases in section two. All self-disclosure accounts, on which these cases would indeed count as failures of free will, are sensitive to the conceptual issue that I will raise in this paper.

But self-disclosure does not only involve authenticity; it also involves regulation. In order to act freely on these accounts, an agent's actions need to be *regulated* in the light of these authentic commitments and cares. Hence, besides meeting authenticity conditions, an agent also needs to meet what I will call the *regulation condition*. I consider this condition to be met when one's motives and desires are in line with one's authentic commitments *or* can be resisted when they diverge from these commitments. In this paper, I will be primarily concerned with what the literature on implicit attitudes can teach us about this specific condition.

One last issue should be sidestepped before I turn to the relevant research on implicit attitudes. Theories of free will that require one's will to be authentic or wholehearted have been criticized for failing to give a satisfactory account of the conditions for responsibility. They are said to be unduly restrictive because we often hold people responsible for behavior that does not express their cares or commitments (see Sripada, 2015). Arguably, the capacity to act on the basis of authentic commitments is, under some definitions, not a necessary condition for moral responsibility. But even if that were true, it changes nothing about the fact that a specific sort of *freedom* is lacking when one fails to

control one's behavior in the light of what one is committed to. I therefore want to bracket issues of moral responsibility here and focus on this conception of freedom for its own sake.

2. Implicit prejudice and a failure to exercise free will

In this section, I will discuss how research on implicit attitudes highlights failures of free will. In the next three sections, I will be particularly concerned with those implicit attitudes that are prejudiced against members of certain stigmatized social groups. Where necessary, I will call these specific implicit attitudes *implicit prejudices*.

A standard definition of implicit attitudes in philosophy is that they are "relatively unconscious and relatively automatic features of prejudiced judgment and social behavior" (Brownstein, 2015a, p. 1). Though this definition of implicit attitudes is common, the use of the term 'unconscious' may give rise to confusion. In light of recent research, it would be incorrect to claim that one is not conscious or aware of implicit attitudes at all. I will say more about this in section 3.1. I will stick to a more minimal definition of an implicit attitude as a relatively automatic feature of biased judgment and social behavior (see De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009).

The Implicit Association Test (IAT) developed by Greenwald, McGhee, Schwartz, and Jordan (1998) reveals the omnipresence of implicit prejudices. The IAT tests one's implicit associations by the speed and level of association made between group concepts and stereotypical evaluative concepts. One can make either stereotype-congruent or stereotype-incongruent associations. If stereotype-congruent associations are made more easily and more often, this indicates an implicit association between a group and a stereotypical property or trait and hence that the participant is subject to implicit stereotypes. The fact that the participant is encouraged to respond quickly ensures that *implicit* associations are measured rather than explicit ones.[2] The vast majority of people participating in this test are shown to harbor implicit prejudices that they explicitly renounce. Among them are strong prejudices against women and Black people. Members of those groups themselves are also implicitly biased against the group to which they belong (Greenwald & Krieger, 2006).

These implicit prejudices have an effect on one's behavior (Bargh, Chen, & Burrows, 1996). They are tendencies or attitudes that function as push and pull factors in action. For example, if Black people are associated with crime, this may give rise to fear and a tendency to avoid or accuse a Black person, and if one associates men with certain professions while failing to associate them with women, one is likely to place more trust in men in these positions. It is easy, then, to imagine how and why these attitudes may govern many interactions outside of the experimental context. Negative associations may give rise to lack of eye-contact with, avoidance of, or lack of trust in or even fear of the object of these associations in all sorts of daily situations.

Different experiments also demonstrate these effects of implicit attitudes on actions. Some of the most striking experiments concern the evaluation of CVs, term papers, or other documents representing one's achievements and skills (Bertrand & Mullainathan, 2004; Goldin & Rouse, 2000; Steinpreis, Anders, & Ritzke, 1999). In these experiments, different employees at universities were asked to evaluate such documents. Each received the same documents but with different names attached to them. Those documents that bore typically White, male names were evaluated as significantly *better* than the *same* CVs which bore typically female, Black, or Arabic names. The implicit stereotypes tested by the IAT reliably predicted the outcomes. Those managers who were more implicitly prejudiced against candidates with female or Arabic names were also less likely to interview an Arabic or female job-applicant. Explicit commitments were also surveyed, and they were discordant with these attitudes (Rooth, 2007). Other research suggests that police officers are more likely to shoot Black persons, that teachers make less eye-contact with female students and give them less time to respond to questions, and that doctors are less likely to fully inform ethnic minorities and poor people (Johnson et al., 2013; Payne, 2001; Sadker & Sadker, 1986).

I contend that these experiments provide us with concrete cases of failures to exercise free will on a self-disclosure account. In these same experiments, explicit attitudes are also measured and people

almost always denounce racism, sexism, or similar attitudes. Implicit prejudices, therefore, seem to conflict with the agent's own commitments. Surely, more research will have to be done in order to find out whether implicit attitudes really *always* fail to express what participants care about and are committed to. People may be self-deceived about what their real cares and commitments are, or these self-reports may be given in order to keep up appearances (see Brownstein, 2015b; Levy, in press). However, it is likely that many participants do care for and are committed to equality and justice and really would denounce the influence of racism, sexism, or other prejudices on their behavior. The case that began this paper is such an example. The employers in the CV-research are other examples. This research suggests that even those people who are really committed to and care for equality can, at times, be guided by undesirable implicit prejudices against stigmatized groups.

This brings us to the conclusion of this section. Research on implicit prejudice suggests that people act on the basis of implicit prejudices *even though* these attitudes fail to align with their authentic commitments to equality. These failures of alignment may here be interpreted broadly as either a lack of identification, endorsement, or a failure to express one's cares. Many, and maybe all, self-disclosure accounts of free will would look upon these cases as failures of free will. Exercising free will means overcoming these failures. Can this be done?

3. Individual resources for control

In this section, I will consider the individual's ability to overcome these failures of free will by reviewing state of the art empirical studies of resources for controlling the effects of implicit prejudice on behavior. I will explain how individual resources for exercising control over implicit attitudes, and specifically implicit stereotypes, do not seem to suffice for *full* control. I aim to be as inclusive as possible by addressing all the types of resources one could put to the task of exercising control over implicit attitudes. While these resources suggest that people have *some* ability to exercise control over their behavior, I argue that research on implicit attitudes indicates that these abilities are significantly limited. It is commonly concluded on the basis of this research that the mechanisms and capacities required for exercising free will are limited. At the end of this section, I will consider this conclusion and explain why it is drawn too hastily.

3.1. Awareness

Lack of awareness of implicit attitudes and their effects on behavior has often been considered one of the main reasons that people lack control over them. But recent studies suggest that one can be aware of the implicit attitudes that one harbors in the same way that one may be aware of one's gut feelings. Additionally, people even seem to be good at predicting the effects that implicit prejudice may have on their behavior (Hahn, Judd, Hirsh, & Blair, 2013). Indirect knowledge provided by empirical research may thereby facilitate an awareness of the influence that implicit prejudice can have on one's behavior.

This being said, I take it that monitoring such influences is still rather complicated. There are good reasons to doubt that the ability to predict the effects of implicit attitudes on one's behavior generalizes to situations and contexts that have not been tested. One problem is that implicit attitudes may influence behavior in unexpected situations. For example, is one required to pay heed to the possible influence of attitudes on behavior when watching television? You may be tempted to say no. But a study has shown that the majority of male television viewers change channels when a discussion program features only women, while they keep watching when the guests are mixed or only male (d'Haenens, 2006). This behavior may be inconsistent with the values and commitments of some of these viewers. But it probably did not occur to them to predict the impact of implicit prejudices when coming home and "just watching TV." One simply cannot *always* foresee the impact that implicit attitudes may have in certain situations. Secondly, even if one is always able to predict the effects of implicit attitudes on one's behavior, one has to actively look out for situations in which they may arise. This implies that one must be vigilant all the time. This exceeds the psychological capacities that people actually have

and may unduly interfere with other tasks that require our attention and control. For these reasons, being aware of the impact of implicit attitudes on one's behavior is taxing, and always being aware of them is just too psychologically demanding. This in turn impedes our ability to control them.

Our ability to exercise control over implicit prejudice is further complicated in other ways. Even in those cases in which one *can* predict the effects of implicit prejudice on action, the crucial question is whether the agent can also resist these effects. I will now discuss the typical resources that an individual may employ to resist diverging motives and show why they are of limited help.

3.2. Willpower

In those cases in which one *can* predict the effect of implicit attitudes on one's behavior, there are still further difficulties when it comes to suppressing these effects. Willpower allows one to resist or override the pull of undesirable motives. This is the kind of strength that one needs if one is to directly regulate the effect of attitudes on one's behavior. But contemporary psychological research suggests that inhibiting the effects of implicit prejudice by means of willpower gives rise to complications.

The evidence for this comes from experiments in which groups of participants required to exercise self-regulatory skills and suppress impulses are shown to perform worse than neutral groups in subsequent tasks that require control (Baumeister, Bratslavsky, Muraven, & Tice, 1998). An example of the kind of test that requires executive control is the Stroop task. In this task, participants are asked to name the font color of words that are the names of (different) colors. Richeson and Shelton (2003), among others, have shown that the more implicitly biased one is (according to the IAT), the worse one performs in these tests after interracial interactions. This suggests that interracial interactions require self-regulatory efforts that compromise the exercise of similar forms of executive self-control in the Stroop task (see also Salvatore & Shelton, 2007; Sripada, Kessler, & Jonides, 2014).[1]

The general suggestion is that the kind of strength of will that is needed to suppress implicit attitudes is often followed by a refractory period. The exact explanation for this is the subject of debate. One explanation for the temporary willpower impairment is the resource model of self-control introduced by Baumeister, Vohs, & Tice, (2007). But replications of this experiment have given rise to criticisms of this account and suggest that alternative explanations should be considered (Hagger & Chatzisarantis, 2016; Sripada et al., 2014). Nonetheless, this research gives voice to the plausible assumption that one simply cannot exercise willpower *all the time*.

Willpower only provides a limited means for controlling one's behavior. A *constant* execution of this type of cognitive control is simply too demanding for human beings. It is therefore unlikely that subjects can always (or even often) rely on executive self-control to correct for implicitly prejudiced responses. This is made more unlikely by the fact that an individual may need these means for different types of cognitively demanding tasks in everyday life.

Even as a limited resource for control, willpower does not always provide a plausible means to exercise control over implicit attitudes. Take the CV case. Even when one is aware of the influence of implicit prejudice on behavior in such situations, it is unclear how exercising willpower could actually help. How is one to stop one's judgments from being biased by means of willpower here? I, for one, wouldn't know how to do that. Though one may stop oneself from crossing the street when seeing a group of Black people or one may force oneself not to avoid eye contact, and so on, in certain other situations a simple exercise of willpower does not help to control the effect of implicit stereotypes. When implicit attitudes influence one's judgments, it seems impossible to control the effect of attitudes by means of willpower. Hence, willpower is not only a limited resource, it is also a limited remedy for the effect of implicit attitudes on one's behavior because not all actions can be regulated by means of willpower. The limited capacity for exercising willpower explains why it is even more difficult to exercise free will when confronted with implicit attitudes. Even when one is aware of the need to exercise control, willpower may be exhausted and hence be an unavailable resource. There are also situations in which willpower is not a resource that can be used for controlling the effects of implicit attitudes. But there are other resources for individual control that should be addressed.

3.3. Implementation intentions

The ability of human beings to internalize behavioral policies and exercise them "unthinkingly" is also constitutive of the ability to live one's life in the light of motives or desires that one would endorse. Does non-reflective control offer a possible means for individuals to prevent the effects of implicit attitudes on their behavior?

Some research has been done on "automatic" or "non-reflective" control over implicit stereotypes by means of implementation intentions. These types of intentions were first conceptualized and tested by Peter Gollwitzer (1999). Implementation intentions are strategies of self-control that can override automatic attitudes by internalizing *contrary* automatic attitudes. These are concrete, goal-directed intentions that are aimed at bringing one's actions in line with one's commitments. An example of an implementation intention is saying to oneself, "At the conference tomorrow, if a woman is talking, then I'll listen." Such intentions can decrease biased response in the situations at which these intentions are directed. For example, they trigger listening as an automatic attitude when a woman is talking and prevent an automatic tendency to refrain from listening when a woman speaks. One thereby does not need to exercise cognitive and deliberative effort all the time, because one conditions oneself to perform such behavior without thinking about it (Madva, 2016; Stewart & Payne, 2008). By means of implementation intentions, one is able to act in a way that is more consistent with reflective commitments. This strategy has been seen as a boost for people's ability to exercise control over their behavior (Gendler, 2014; Holroyd, 2012; Stewart & Payne, 2008).

It is not clear that implementation intentions will give us lasting and significant control. They have only been shown to apply to those specific situations that one primes oneself to respond to at a particular time. They are effective because, once formed, the specific situation automatically triggers the intended response. This means one has to train and prepare oneself for many different situations. Apart from the intention to listen when a woman talks at a conference, one should also internalize self-governing policies like, "I should not get disproportionally angry at a Black person bumping into me," "I should not change the channel from discussion programs featuring women," "I should not expect a doctor or a professor to be a man," "I should not expect people with an accent from the countryside to be less intelligent," and so on. The list is endless.

This first means that the degree of self-training that would be required to counter the effect of certain automatic attitudes is impossibly high. One can imagine that most people do not even value their own autonomy enough to go through the trouble of forming implementation intentions for so many different situations. A second and related problem is that it is simply impossible to take *unexpected* situations into account. If one does not know *which* situations demand an implementation intention, one cannot form such an intention. This brings us back to the problem of detection. One often does not *know* which situations are likely to trigger an undesirable automatic attitude. This being said, implementation intentions do make for an important subset of actions that an individual can control.

3.4. Triggers

One other way in which individuals can exercise control is by manipulating the environment. Such manipulations may be instrumental to exercising free will. For example, if you are typically late for work because you cannot stop reading your novel during breakfast, an obvious solution is to hide your novel and instead read the paper. Similarly, if you have a tendency to buy and eat cake when you see it, you can avoid walking past bakeries, and so on. By simply avoiding or altering certain triggers that are likely to lead to a failure to exercise free will, one can prevent failures of free will.

Ironically, one may prevent the effects of implicit stereotypes on behavior by simply avoiding the people against whom one is biased. The triggers that give rise to biased behavior are, after all, members of socially stigmatized groups. But obviously these means defeat the very aim. One does not want implicit stereotypes to govern one's actions because one does not want to enact stigmas and reinforce stereotypes. But by avoiding people one is likely to be doing exactly that. Also, the desire to treat

people equally isn't realized by not treating them at all. Therefore, at first sight, avoiding the triggers for implicit stereotypes does not seem to be a feasible solution to one's inability to exercise control over the effects of implicit stereotypes on one's behavior.

Nevertheless, certain ways of avoiding triggers can sometimes offer local solutions to biased behavior. Anonymous review is a method of this kind. By reviewing work anonymously, one avoids the triggers that may give rise to biased judgments, and these means encourage a more just and equal treatment of others. This is a strategy for exercising free will because in doing so one is making sure that one's judgments are not based on motives or desires that one would not endorse. One does so by *avoiding* the possibility that stereotypical names will trigger implicit stereotypes which consequently influence judgment. Similar strategies employed to avoid the triggering of bias may sometimes offer feasible ways to secure one's ability to exercise free will.

However, these strategies can again only be employed when one is able to predict when and where they are necessary. As has already been discussed, this cannot always be done, and even when it can be done, it places quite a burden on the subject. Additionally, even if effective strategies can be envisioned, these cannot always be executed by an individual alone. They sometimes require institutional changes. Large scale anonymous reviewing, for example, requires assistance in removing names and distributing feedback. If such assistance is not available, and can't be made available, one cannot employ a strategy of anonymous review.

3.5. The hasty conclusion

The research discussed suggests that even people who are sincerely committed to equality lack, to a significant extent, the individual resources to resist the effects of implicit attitudes on their actions. On a self-disclosure account, exercising free will means overcoming such failures. This section began with the question of whether one can overcome the failures of free will discussed in section two. One may be tempted to say that the answer to this question has now been found. Although there is some limited space for regulation, people are, to a significant extent, unable to exercise free will in light of implicit attitudes. Consequently, it could be concluded that free will is a limited capacity. In the literature, this conclusion can indeed be found. It is argued for by Neil Levy and is also common in the situationist literature (e.g., Levy, in press; Nelkin, 2005).

But this conclusion is too hastily drawn. Note that only *individual* capacities for control have been shown to be limited. If one concludes that free will is a limited capacity purely on the basis of these findings, one implicitly assumes that free will is *by definition* an individual capacity. Because *only* individual capacities have been shown to be limited, we can only validly conclude that free will is a limited capacity when free will is an individual capacity. But not all theories of free will are committed to the claim that free will is an individual capacity. Self-disclosure accounts of free will in particular are not explicitly committed to this. Let me spell out the "hasty conclusion" as it would apply to the self-disclosure accounts of free will discussed in section one:

P1. Free will requires coherence between one's motives and one's authentic commitments or the ability to resist diverging motives and desires.

P2. Research on implicit attitudes suggests the individual capacity to resist diverging motives is limited.

C. Free will is a limited capacity.

In premise two, it is only claimed that research on implicit attitudes suggests that the *individual* capacity to resist diverging inclinations is limited. So, really all that can be concluded here is that *individual* abilities for exercising free will are limited. This is altogether different from the hasty conclusion that the capacity to exercise free will is limited. It is different because the capacity to exercise free will *need not* be conceptualized as individual on a self-disclosure account. It is nowhere mentioned in these theories that the abilities required for disclosing oneself are necessarily purely individual. The only thing that is definitive of these abilities, whatever they are, is that they should allow the agents to express their authentic commitments in their actions.

If there were non-individual abilities that would allow for the exercise of self-regulation, this conclusion would be false. In other words, if self-regulation, and therefore free will, were a socially embedded or extended capacity that *could* be exercised such as to control for the effects of implicit prejudice, the conclusion that free will is a limited capacity would be false.

I will now continue to ask what the literature on implicit attitudes can tell us about the ability to exercise free will. I do so by asking what *would* be required for exercising free will in the case of implicit prejudice. As it turns out, the answer to this question indeed shows the hasty conclusion to be unwarranted.

4. The social context, associative learning and free will

In this section, I will first explain how research suggests that implicit attitudes are socially constituted and maintained, and second I will discuss how this implies that an individual cannot significantly modulate implicit prejudice all by oneself. These two explanations show that failures to self-regulate in the light of implicit prejudice are due to the social structures we grow up in and exist in. They provide a basis for answering the question of what would be required to overcome failures to exercise free will.

4.1. Associative learning and the constitution of bias

The general contention in social psychology is that implicit behavioral attitudes and processes develop through different forms of associative learning or conditioning (Mandelbaum, 2015). In this section, I will focus on those specific attitudes that are stigmatizing—implicit prejudices. Bryce Huebner discusses the relation between these learning processes and implicit prejudice in detail (Huebner, 2016). Different types of learning processes seem to be involved in the constitution of implicit prejudice.

One way in which implicit prejudices are likely to come about is by Pavlovian association. Pavlovian association is a quick and stable association that builds on innate reflexes and responses. It yields attitudes grounded in basic emotions like fear, disgust, and sexual lust. Dunham, Baron, and Banaji (2008) have argued that implicit attitudes appear quite early in life and remain stable over the lifespan. Some very primary and instinctive attitudes produce the kinds of mechanisms that later in life take the shape of implicit attitudes. Toddlers seem to prefer the faces of their primary caregivers, their native language, women, and racial in-group members. These preferences are not innate but learned. Babies whose primary caregivers are male or racial out-group members do not have these preferences. Such attitudes are clearly not stigmatizing in nature, but they are likely to play a role in the development of implicit prejudice (Dunham et al., 2008).

Later (but not much later) in age, the contour of implicit prejudice begins to take shape. Research based in the U. S. shows that White children of six already show strong in-group preferences and associate negatively with out-groups (Dunham et al., 2008). The same research involving Hispanic children shows *no* in-group preference when compared with White Americans, while there *is* an in-group preference when comparisons are made between Hispanics and Black Americans. Black children of five show no in-group preference at all (Dunham et al., 2008). This suggests implicit attitudes that are related to stigmas and stereotypes, and they seem to be rapidly internalized at a very young age. A related suggestion from developmental psychology is that in-group preferences and the ability to rapidly evaluate groups are essential to survival and are probably evolved mechanisms (Cosmides, Tooby, & Kurzban, 2007; Kurzban, Tooby, & Cosmides, 2001). These basic Pavlovian mechanisms of group evaluation are also triggered by existing stigmas about groups.

Implicit associations are probably also realized through slower forms of conditioning that allow for the revision of behavioral policies in order to increase the likelihood of bringing about certain "valuable" outcomes. Such forms of reinforcement learning are also typically unconscious but are slightly more complex than Pavlovian models of learning. In these more complex forms of associative learning, one learns from the consequences of one's own behavior by means of a much larger variety of subtle rewards or punishments (Huebner, 2016).

These processes are also responsible for implicit attitudes because they *reinforce* them when they conform to predicted positive outcomes (i.e., a biological reward). While Pavlovian systems respond to an immediate "threat-danger" cue by avoidance responses, these associative processes also develop in response to subtle norms and social instruction (Klucharev, Hytönen, Rijpkema, Smidts, & Fernández, 2009; Klucharev, Munneke, Smidts, & Fernández, 2011). Neuroscientific research has indicated that conformity to social norms indeed corresponds to such a positive biological "reward" in this sense (Klucharev et al., 2009).

Huebner (2016) concludes that in real-world environments, people unconsciously learn from and respond to explicit and implicit warnings and normative associations about certain groups. He observes that these forms of reinforcement are omnipresent in stereotypical associations. Examples include advertisement and general media representations of demographic groups or implicit verbal and behavioral instructions about the threats and dangers posed by stigmatized groups and the situations in which such danger is likely to arise. These are the kinds of cues that implicit learning responds to. The relationship between social norms and cues and implicit social learning explains how the social context figures into the constitution and maintenance of implicit attitudes.

4.2. Modulating implicit attitudes

Can an individual alter the stigmatizing character of his or her implicit attitudes? A reflective judgment that an association is not warranted, or even undesirable, does not in itself change the associations one has. But a promising alternative is to get rid of earlier associations through modified forms of associative learning. Getting rid of implicit prejudice then requires changing the environmental cues through which these prejudices are realized and reinforced. There are two ways in which this can be done. First, prejudiced attitudes change if the actions that result from them are repeatedly met with negative social and environmental feedback. These prejudices also change when counter-stereotypical associations are repeatedly positively reinforced.

Counter-stereotypical images and associations have been shown to decrease biased responses in math tests for women. Women that are subtly trained to associate women with good performance at math-related activities actually perform better than those who are not (Madva, 2015). It has been reported that people who purposefully and continuously confront themselves with counter-stereotypical images and associations have decreased implicit prejudice for at least eight weeks (Devine, Forscher, Austin, & Cox, 2012). This research is promising because it seems that environmental manipulations intended to break up or oppose biased associations can be successful.

But there are good reasons to be skeptical about the overall long-term and large-scale effects of these types of strategies. In order to be completely rid of implicit stereotypes, one needs to *remain* subject to other forms of associative learning, ideally from a young age. Associative structures quickly consolidate, and once they do they will always be latent and easily recover when one is confronted with certain associations. If one really wants to be rid of implicit stereotypes, one needs to be wholly differently conditioned.

This, I believe, is not something that one can do all by oneself. Our meaningful and instructive social environment is something we share and create as a community. Therefore, only a community can change the environmental associations and instructions that constitute implicit stereotypes. An individual alone can't possibly realize an environment that secures negative reinforcement of stigma and positive reinforcement of de-stigmatizing associations. Only as a community can we ensure an environment characterized by such forms of "egalitarian conditioning."

It was argued in section three that an individual cannot fully resist the effects of implicit prejudice on her behavior. Because an individual also cannot modulate implicit prejudices all by herself, it can now be concluded that, to an extent, an individual really *lacks* the capacity for self-regulation. And secondly, it may now be concluded that this inability is due to the specific social relations in which one stands, because implicit attitudes toward members of socially stigmatized groups are constituted and reinforced through the stigmatizing associations that are expressed in a community. Taken together,

these claims provide the basis for answering the question of what is required for the exercise of free will. They suggest that a community *can* secure forms of reinforcement that lead to the modulation of undesirable prejudiced attitudes. And to the extent that a community can do so, the ability to self-regulate should be conceptualized as social. As a consequence, the capacity for free will should be conceptualized as partly social as well. In the next section, I discuss this answer in more detail and situate it in the existing literature on social freedom.

5. The social capacity for free will

The limitations on the individual capacity to exercise free will and the discussed types of implicit social learning form the basis for my argument for the social re-conceptualization of free will. I will situate this argument in the debate on relational autonomy. Relational autonomy is the umbrella term for theories of autonomy that aim to re-conceptualize self-government in the light of the social embeddedness of human beings (Mackenzie & Stoljar, 2000). This debate is relevant for two reasons.

First, the accounts of autonomy that are central to this debate typically imply self-disclosure accounts of free will (Anderson & Honneth, 2005; Christman, 2004; Stoljar, 2013). My conclusions about self-disclosure accounts are directly relevant to relational theories of autonomy that imply a form of self-government that consists in the ability to express one's authentic self in one's actions.

Second, this debate helps to classify the specific *kind* of re-conceptualization I propose. It can hardly be denied that social relations can both obstruct and foster free will. But what is contested is what the exact conceptual relation is between the social context and self-government. The relational autonomy debate helps to provide a taxonomy of the different ways in which social conditions relate to freedom. By means of this taxonomy, I will analyze the specific *kind* of social relation that implicit-bias research exposes. In conclusion, I will discuss an objection to the proposed conceptualization of free will and show that it has a nasty bullet to bite.

5.1. A taxonomy of social conditions

There are roughly three different ways in which social relations are thought to relate to free will. Many views have emphasized the necessary social conditions for the *development* of self-governance (e.g., Baier, 1985; Meyers, 1989). They focus on the social conditions involved in the genesis of abilities for authentic self-rule. That the ability to exercise free will is socially embedded in this sense is a rather uncontroversial claim. Learning a language and being nurtured are social undertakings, and one can hardly develop the necessary abilities for self-rule without them. On such a relational account, free will can still be analyzed in terms of individual capacities, but the capacity necessarily develops by means of specific forms of social interaction.

Besides focusing on the development of free will, one may also consider the sort of social context within which free will can flourish. Social oppression, for example, is typically seen as a social condition that threatens autonomy, while good decision-making communities and loving relationships foster and facilitate autonomy. These theorists argue that self-government can be impacted or affected by certain social conditions, but does not stand in a necessary or conceptual relation to these social conditions. The capacities for free will are themselves individual and independent. The accounts of relational autonomy on which autonomy can be analyzed in terms of individual abilities, without reference to social relations, are often categorized as "causal" theories.

On stronger accounts, the relationship between social conditions and the capacity for free will is taken to be of a necessary and conceptual character. These accounts are named 'constitutive' theories of autonomy. When social relations are constitutive of autonomy, they are necessary for self-government and are *given with* the definition of autonomy (Christman, 2004; Mackenzie & Stoljar, 2000; Stoljar, 2013). So, for example, if a decision-community is argued to be constitutive of autonomy, one cannot be autonomous without it. Marina Oshana, for example, has argued that the access to significant social resources is a constitutive condition of autonomy. This implies that *by definition* one cannot be

autonomous in an oppressive society (Oshana, 1998; for discussion see Christman, 2009). Recognition theories of autonomy sometimes seem to imply that forms of social recognition are constitutive elements of autonomy as well (Christman, 2009).

5.2. A social re-conceptualization of self-regulation

This taxonomy helps to identify my position in this debate. The last category is the one that fits my argument. Before explaining this, the lesson learned from the discussion on implicit attitudes and social learning merits emphasis.

In order to overcome failures of free will due to implicit prejudice, one needs to resist or be rid of those implicit attitudes that are stigmatizing and in conflict with one's egalitarian commitments. This, in turn, requires standing in the right sort of community—a community that does *not* reinforce stigmatizing associations and instead positively reinforces non-prejudiced evaluations of social groups. In addition, there are most likely other types of implicit attitudes that already help the agent to act in ways that align with the cares and commitments that she has. Also, research suggests that implicit attitudes can easily become endorsed and accepted by the agent (Levy, in press). It is crucial to see that those implicit attitudes that *align* with one's cares and commitments cannot be detached from the social context in which they are reinforced. Our motives and desires, whether in conflict or in harmony with cares and commitments, are maintained and constituted by the social context *whenever* they involve an implicit attitude.

I can now turn to the social re-conceptualization of free will and discuss how it fits into the established taxonomy. Recall the regulation condition: free will requires the regulation of one's behavior in light of one's authentic cares and commitments. The extent to which one meets this condition is dependent on how one's implicit attitudes align or conflict with one's cares and commitments. When a social situation reinforces *diverging* implicit attitudes, this is always a limitation to self-regulation because, as I have argued in section three, people have limited abilities to resist acting on the basis of diverging implicit attitudes and because, as I have argued in section four, people are not individually able to properly modulate these attitudes. When, on the other hand, these attitudes align with one's cares and commitments, the ability to regulate oneself is constituted or enabled by the social situation in which one stands.

The regulation condition, therefore, cannot be detached from the social situation in which one stands. It follows that one's level of free will cannot be detached from these social situations. Relations that reinforce desirable implicit attitudes are necessary for free will. Having free will *means*, to some extent, standing in the right sort of relations to others. This is because certain aspects of the ability to exercise free will simply cannot be explained without making reference to the social relations that reinforce implicit attitudes. This account therefore seems to fit into the last category of the taxonomy. It is a constitutive relational theory.

But it does not fit into the taxonomy very neatly. I have only argued that free will is *co*-constituted by the social. Social relations are not completely essential to free will. They are only necessary for free will *when* our motives are comprised of implicit attitudes that escape our control, which they often are. In section three, it became clear that there is limited space for individual control when resisting the effects of implicit prejudice. But implicit attitudes are not the only sorts of motives that people have. Other types of motives may be subject to individual control, whereas implicit attitudes are not. Therefore, on the basis of my argument, it can only be concluded that free will is *partly* constituted by the social. It is a further (empirical) question to what extent, and for which motives and attitudes, our will can be said to be socially co-constituted in this sense. But research on implicit prejudice already suggests that the co-constitution is significant.

There is one more caveat. It is specifically the *regulation condition* that should be regarded as co-constituted by our social environment. The ability to *form* authentic commitments and cares has not been addressed in this paper. This is atypical. Most theories are concerned with the social conditions for *authenticity* (e.g., Christman, 2009; Taylor, 1985). I have no quarrel with Christman's claim that one

can form authentic commitments and cares when living under socially oppressive conditions (2009). But I am arguing that the social environment is always involved in one's ability to *act* on the basis of authentic commitments and cares. It is this specific sub-capacity of free will that I have argued is socially embedded or co-constituted.

The vigilant reader may have noticed that I claim capacities for free will to be *either* socially constituted *or* embedded. In the relational autonomy debate, not much weight is placed on the distinction between the two. Necessary conditions, background conditions, and constitutive conditions for autonomy are not clearly distinguished in the literature. (e.g., Christman, 2004; Mackenzie & Stoljar, 2000; Stoljar, 2013). Arguably, there is a difference between these conditions that deserves to be fleshed out. Those interested in ontological questions about human capacities place great stress on distinguishing constitutive conditions on the one hand from scaffolding or supporting conditions on the other, most prominently theorists working on the extended mind and will. In this paper, I do not mean or need to takes sides. What matters for my argument is that specific social conditions are *necessary* conditions for the capacity to exercise free will and should therefore be made reference to in the *definition* of free will.

5.3. A possible objection

One may reasonably object that one *could* still maintain that free will is a purely individual capacity in the light of these studies. But because self-disclosure accounts do not claim that free will is necessarily exercised through individual capacities, one is no longer defending a self-disclosure account of free will were one to do so. And there is reason not to do so. If one wants to maintain that free will is by definition only exercised through independent, individual capacities, one has a bullet to bite.

Those who want to maintain that free will is a purely *individual* capacity would have to maintain that our will is not free whenever we act on the basis of implicit attitudes, even when they line up with our authentic cares and commitments. This is because the ability to act freely, in those instances, is not the function of an individual exercise of control. I take it that this is a counter-intuitive conclusion. One would, for example, have to say that one's will is not free when one avoids dark and smelly alleys if this response is explained by implicit attitudes, e*ven if* this is perfectly consistent with one's authentic cares and commitments. Something similar may sometimes have to be said about being guided by preferences for family members and close friends. If implicit social reinforcement motivates one to behave beneficently toward a loved one, this action would not be an instance of free will even if it aligns with an authentic commitment. If one is unwilling to bite this bullet and wants to maintain that one's will is free *whenever* one acts on the basis of motives or desires that do not conflict with one's authentic cares and commitments, as self-disclosure views typically do, then one may want to opt for the alternative conclusion set out in this paper.[3]

Notes

1. In calling these notions 'self-disclosure' notions of free will, I follow Watson (1996).
2. For more information on the test, see also implicit.harvard.edu.
3. I want to thank Jan Bransen, Julian Kiverstein, Beate Roessler, Marc Slors, and two anonymous referees for their helpful feedback on earlier drafts of this paper.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by The Netherlands Organisation for Scientific Research' [grant number 36020360].

ORCiD

Daphne Brandenburg http://orcid.org/0000-0002-6942-6902

References

Anderson, J. H., & Honneth, A. (2005). Autonomy, vulnerability, recognition, and justice. In A. Honneth & J. H. Anderson (Eds.), *Autonomy and the challenges to liberalism: New essays* (pp. 127–149). New York, NY: Cambridge University Press.

Baier, A. (1985). *Postures of the mind: Essays on mind and morals*. Minneapolis, MN: University of Minnesota Press.

Bargh, J., Chen, M., & Burrows, L. (1996). The automaticity of social behavior: Direct effects of trait concept and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230–244.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74*, 1252–1265.

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science, 16*, 351–355.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*, 991–1013.

Bratman, M. (2007). *Structures of agency: Essays*. Oxford: Oxford University Press.

Brownstein, M. (2015a). Implicit bias. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2015 ed.). Retrieved from http://plato.stanford.edu/entries/implicit-bias/

Brownstein, M. (2015b). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology, 6*, 1–22.

Christman, J. (2004). Relational autonomy, liberal individualism, and the social constitution of selves. *Philosophical Studies, 117*, 143–164.

Christman, J. (2009). *The politics of persons, individual autonomy and socio-historical selves*. Cambridge: Cambridge University Press.

Cosmides, L., Tooby, J., & Kurzban, R. (2007). Perceptions of race. *Trends in Cognitive Science, 7*, 173–179.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48*, 1267–1278.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*, 347–368.

d'Haenens, L. (2006). Media as managers of diversity. In L. d'Haenens, M. Hooghe, H. Gezduci, & D. Vanheule (Eds.), *'New' citizens, new policies: Developments in diversity policy in Canada and Flanders* (pp. 137–157). Gent: Academia Press.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Science, 12*, 248–253.

Frankfurt, H. (1988). *The importance of what we care about: Philosophical essays*. Cambridge: Cambridge University Press.

Gendler, T. S. (2014). I—the third horse: On unendorsed association and human behavior. *Aristotelian Society Supplementary, 88*, 185–218.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *The American Economic Review, 90*, 715–741.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*, 493–503.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Greenwald, A., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review, 94*, 945–967.

Hagger, M. S., & Chatzisarantis, N. L. D. (2016). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*, 546–573.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2013). Awareness of implicit attitudes. *Journal of Experimental Psychology, 143*, 1369–1392.

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy, 43*, 274–306.

Huebner, Bryce (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy*, Vol. 1 (pp. 47–79). Oxford: Oxford University Press.

Johnson, T. J., Weaver, M. D., Borrero, S., Davis, E. M., Myaskovsky, L., Zuckerbraun, N. S., & Kraemer, K. L. (2013). Association of race and ethnicity with management of abdominal pain in the emergency department. *Pediatrics, 132*, e851–e858.

Klucharev, V., Munneke, M. A. M., Smidts, A., & Fernández, G. (2011). Downregulation of the posterior medial frontal cortex prevents social conformity. *Journal of Neuroscience, 31*, 11934–11940.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron, 61*, 140–151.

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *PNAS, 98*, 15387–15392.
Levy, N. (in press). Implicit bias and moral responsibility: Probing the data. *Philosophy and Phenomenological Research*.
Mackenzie, C., & Stoljar, S. (Eds.). (2000). *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. New York, NY: Oxford University Press.
Madva, A. (2015). Biased against de-biasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice. Unpublished manuscript.
Madva, A. (2016). Virtue, social knowledge, and implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy*, Vol. 1 (pp. 191–215). Oxford: Oxford University Press.
Mandelbaum, E. (2015). Associationist theories of thought. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Fall 2015 ed.). Retrieved from http://plato.stanford.edu/entries/associationist-thought/
Meyers, D. T. (1989). *Self, society, and personal choice*. New York, NY: Columbia University Press.
Nelkin, D. (2005). Freedom, responsibility and the challenge of situationism. *Midwest Studies in Philosophy, 29*, 181–206.
Oshana, M. (1998). Personal autonomy and society. *Journal of Social Philosophy, 29*, 81–102.
Payne, K. B. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181–192.
Richeson, J. A., & Shelton, N. J. (2003). When prejudice does not pay: Effects of interracial contact on executive function. *Psychological Science, 14*, 287–290.
Rooth, D. (2007). Implicit discrimination in hiring: Real world evidence. *IZA Discussion Paper No. 2764*. Retrieved from http://ssrn.com/abstract=984432
Sadker, M., & Sadker, D. (1986). Sexism in the classroom: From grade school to graduate school. *Phi Delta Kappan, 67*, 512–515.
Salvatore, J., & Shelton, N. J. (2007). Cognitive costs to exposure to racial prejudice. *Psychological Science, 18*, 810–815.
Sripada, Chandra (2015). Self-expression: A deep-self theory of moral responsibility. *Philosophical Studies, 173*, 1203–1232.
Sripada, C., Kessler, D., & Jonides, J. (2014). Methylphenidate blocks effort-induced depletion of regulatory control in healthy volunteers. *Psychological Science, 25*, 1227–1234.
Steinpreis, R., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles, 41*, 509–528.
Stewart, B., & Payne, B. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin, 34*, 1332–1345.
Stoljar, N. (2013). Feminist perspectives on autonomy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2015 ed.). Retrieved from http://plato.stanford.edu/entries/feminism-autonomy/
Taylor, C. (1985). *Philosophy and the human sciences: Philosophical papers 2*. Cambridge: Cambridge University Press.
Watson, Gary (1996). *Two faces of responsibility. Philosophical Topics, 24*, 227–248.
Watson, Gary (1975). Free agency. *Journal of Philosophy, 72*, 205–220.